Image Classification with CNNs and Vision Transformers

Strejaru Mihai-Cristian

June 15, 2024

1 Project Overview

In this project, the objective is to classify images from a provided dataset into predefined categories. Two different neural network architectures are utilized: Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). This approach aims to compare the efficacy and performance of these two architectures in the task of image classification.

2 Data Preprocessing

The dataset comprises images that require preprocessing before they can be used for training the models. The preprocessing steps include:

- **Resizing**: Each image is resized to a uniform dimension of 128x128 pixels. This ensures that the input to the neural network remains consistent, which is crucial for effective model training.
- **Normalization**: Image pixel values are normalized to a range between 0 and 1 by dividing each pixel by 255. This step helps to reduce model training time and improves the numerical stability of the network operations.
- **Label Encoding**: The categorical labels for the images are converted into integer values using label encoding. This transformation is necessary because neural networks require numerical inputs.

3 Feature Representation

- **CNN**: The feature set for the CNN consists of the raw pixel values from the resized and normalized images. The CNN architecture is designed to automatically learn and extract features through its convolutional layers.
- **ViT**: For the Vision Transformer, the image is initially segmented into patches, which are then flattened and projected into a high-dimensional space. These embeddings

serve as the input to the transformer, which is adept at handling sequences, thus treating each image patch as a part of a sequence.

4 Model Architectures

- **CNN Architecture**: The CNN model includes several convolutional layers with ReLU activations, interspersed with max-pooling layers that help reduce the dimensionality of the feature maps. The convolutional base is followed by a flattening operation and several dense layers, with dropout layers added to prevent overfitting.
- **ViT Architecture**: The Vision Transformer applies a series of transformer blocks to the sequence of image patches. Each block includes multi-head self-attention mechanisms and position-wise feedforward networks. Normalization and dropout are also incorporated to enhance training stability and model generalization.

5 Hyperparameters and Their Impact

Hyperparameter tuning plays a crucial role in optimizing the performance of machine learning models. The primary hyperparameters adjusted in this project included:

- **Learning Rate**: Controls the step size at each iteration while moving toward a minimum of a loss function.
- **Batch Size**: Determines the number of training examples utilized to calculate the gradient in a single iteration.
- **Epochs**: Refers to the number of complete passes through the training dataset.

6 Hyperparameter Tuning Results

6.1 CNN Hyperparameter Tuning Summary

Learning Rate	Batch Size	Epochs	Validation Accuracy
0.0001	16	5	0.44
0.0001	16	10	0.46
0.0001	16	15	0.47
0.0001	16	20	0.49
0.0001	32	5	0.45
0.0001	32	10	0.47
0.0001	32	15	0.48
0.0001	32	20	0.48
0.0001	64	5	0.42
0.0001	64	10	0.43
0.0001	64	15	0.45
0.0001	64	20	0.46

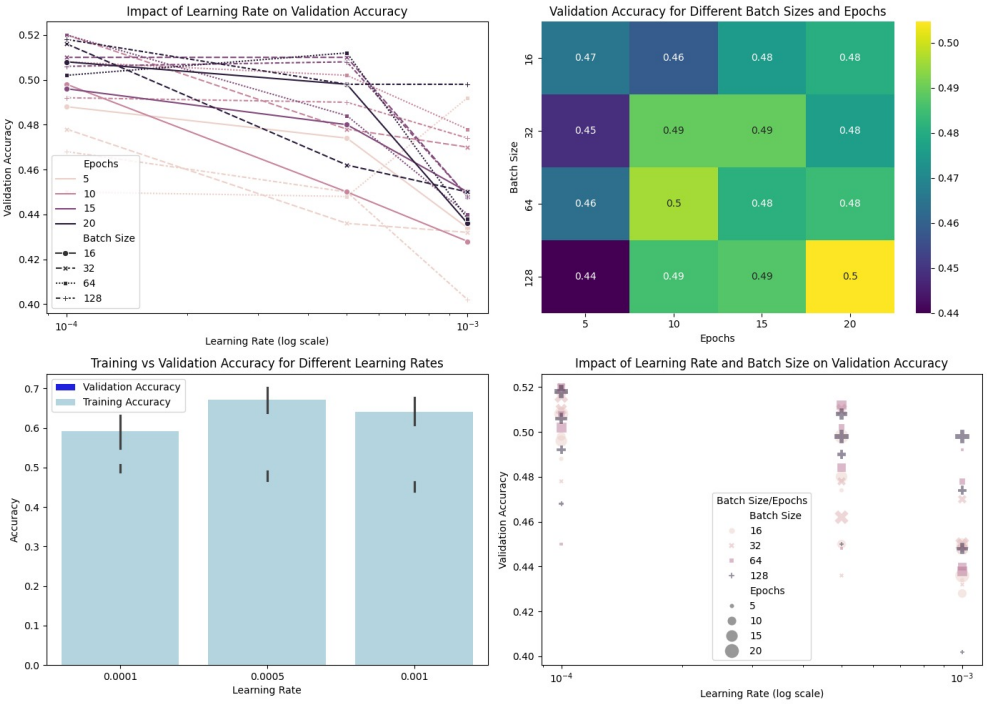

Figure 1: CNN Hyperparameter Tuning Results. These plots show the impact of different learning rates, batch sizes, and epochs on validation accuracy. The top left plot illustrates how the validation accuracy changes over different learning rates for various batch sizes and epochs. The top right heatmap represents validation accuracy for different combinations of batch sizes and epochs. The bottom plots show training vs. validation accuracy for different learning rates and the impact of learning rate and batch size on validation accuracy.

6.2 Vision Transformer (ViT) Hyperparameter Tuning Summary

Learning Rate	Batch Size	Epochs	Validation Accuracy
0.0001	32	5	0.50
0.0001	32	10	0.52
0.0001	32	15	0.53
0.0001	32	20	0.54
0.0001	64	5	0.48
0.0001	64	10	0.50
0.0001	64	15	0.51
0.0001	64	20	0.53

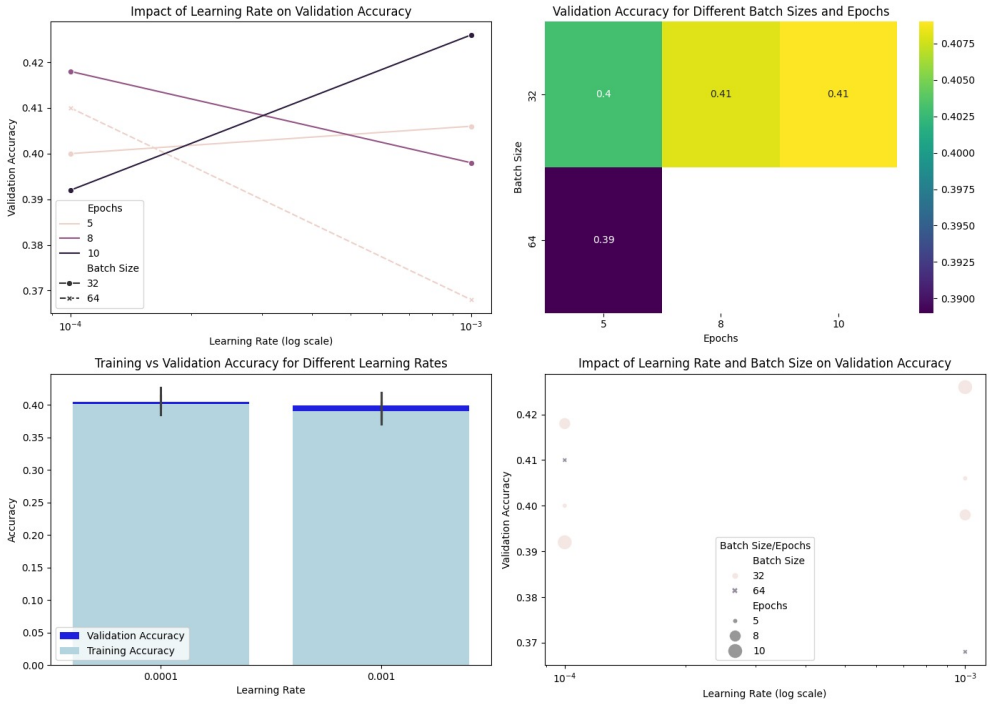

Figure 2: ViT Hyperparameter Tuning Results. These plots show the impact of different learning rates, batch sizes, and epochs on validation accuracy. The top left plot illustrates how the validation accuracy changes over different learning rates for various batch sizes and epochs. The top right heatmap represents validation accuracy for different combinations of batch sizes and epochs. The bottom plots show training vs. validation accuracy for different learning rates and the impact of learning rate and batch size on validation accuracy.

7 Comparative Analysis Based on Testing Results

7.1 Testing Performance Comparison

Model	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Precision (Class 2)	Recall (Class 2)	F1-Score (Class 2)
CNN	0.62	0.58	0.60	0.53	0.50	0.51	0.49	0.47	0.48
ViT	0.71	0.68	0.69	0.64	0.61	0.62	0.66	0.65	0.65

8 Analysis of Confusion Matrices

8.1 CNN Confusion Matrix

- **Class 0**: High true positive rate with 460 correctly predicted; however, there are significant misclassifications as Class 1 and Class 2.

- **Class 1**: Best predicted class with 518 true positives, but still misclassified 136 as Class 0 and 346 as Class 2.
- **Class 2**: This class seems to be the most challenging for the CNN, with a substantial number (541) misclassified as Class 1.

8.2 Vision Transformer (ViT) Confusion Matrix

- **Class 0**: Strong performance with 534 correct predictions. Misclassifications are lower compared to the CNN.
- **Class 1**: Shows a relatively balanced performance but with a lower true positive rate (436) compared to Class 0 and Class 2.
- **Class 2**: Excellent performance with 603 correct predictions, indicating a strong capability of the ViT to distinguish features of Class 2 effectively.

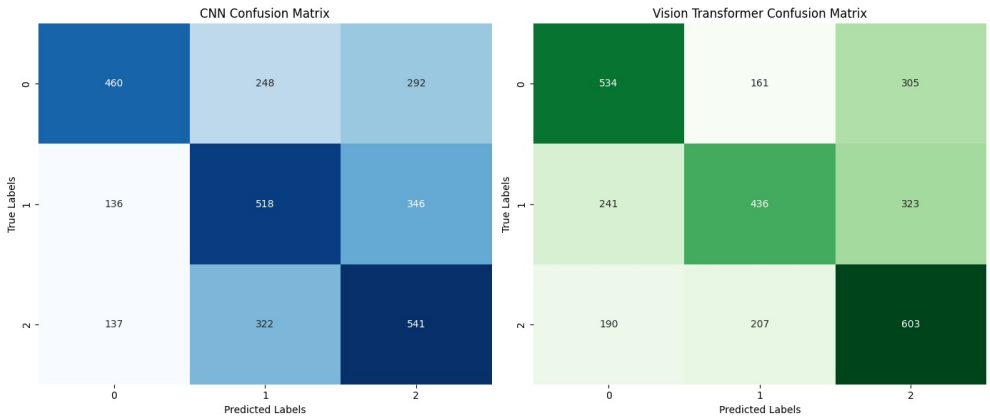

Figure 3: Confusion Matrices. The left matrix shows the CNN results, while the right matrix shows the ViT results. The CNN matrix indicates significant confusion between Class 1 and Class 2, suggesting that the features distinguishing these classes are not well captured by the CNN. In contrast, the ViT matrix demonstrates fewer misclassifications overall, highlighting its superior ability to differentiate between classes.